

DOCUMENT RESUME

ED 309 190

TM 013 661

AUTHOR van der Linden, Wim J.; Eggen, Theo J. H. M.
TITLE An Empirical Bayes Approach to Item Banking. Project Psychometric Aspects of Item Banking No. 6. Research Report 86-6.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE 86
NOTE 37p.
AVAILABLE FROM Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Research/Technical (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bayesian Statistics; Foreign Countries; *Item Banks; Latent Trait Theory; Mathematical Models; Statistical Analysis; Test Construction
IDENTIFIERS Calibration; *Empirical Bayes Estimation; Item Parameters; Optimization; Paired Comparisons; *Rasch Model

ABSTRACT

A procedure for the sequential optimization of the calibration of an item bank is given. The procedure is based on an empirical Bayes approach to a reformulation of the Rasch model as a model for paired comparisons between the difficulties of test items in which ties are allowed to occur. First, it is indicated how a paired-comparisons design deals with the usual incompleteness of calibration data and how the item parameters can be estimated using this design. Next, the procedure for a sequential level of single subjects responding to pairs of items and then for groups of items and subjects of any size. The paper concludes with a discussion of the choice of the first priors in the procedure and the problems involved in its generalization to other item response models. (One graph illustrates a paired-comparisons design for item banking.)
.(Author)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED309190

An Empirical Bayes Approach to Item Banking

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

J. NELISSEN

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

**Research
Report**

86-6

Wim J. van der Linden
Theo J.H.M. Eggen

Department of Education

Division of
Educational Measurement
and Data Analysis

University
of
Twente

Tm013661

Project Psychometric Aspects of Item Banking No. 6

Colophon

Typing : A. Burchartz, L. Padberg
Cover design : M. Driessen, AV-section, University of Twente
Printed by : Central Reproduction Department, University of Twente

An Empirical Bayes Approach to Item Banking

Wim J. van der Linden

Theo J.H.M. Eggen ¹⁾

To appear in Applied Psychological Measurement, 1986, 10

Abstract

A procedure for the sequential optimization of the calibration of an item bank is given. The procedure is based on an empirical Bayes approach to a reformulation of the Rasch model as a model for paired comparisons between the difficulties of test items in which ties are allowed to occur. First, it is indicated how a paired-comparisons design deals with the usual incompleteness of calibration data and how the item parameters can be estimated using this design. Next, the procedure for a sequential optimization of the item parameter estimators is given, first at the level of single subjects responding to pairs of items and then for groups of item and subjects of any size. The paper concludes with a discussion of the choice of the first priors in the procedure and the problems involved in its generalization to other item response models.

Keywords: Item Banking, Item Response Theory, Rasch Model, Paired-Comparisons with Ties, Empirical Bayes Approach.

An Empirical Bayes Approach to Item Banking

An innovative idea produced by modern test theory is the notion of a calibrated item bank. A calibrated item bank is a large collection of test items all measuring the same ability or domain of knowledge that is stored in a computer together with empirical estimates of its item parameters. The item parameters are defined by an appropriate response model fitting the responses of the subjects to the items in the bank. The use of a calibrated item bank instead of standardized tests has two important advantages. The first advantage is the introduction of flexibility in the practice of testing in education and psychology. Using a calibrated item bank, test scores may be obtained from any selection of items from the bank and nevertheless measure the ability on the same scale. This opens the possibility to construct tests from the bank solely on the basis of practical considerations. The second advantage of item banking is efficient use of response data. Any new set of data, even the responses of a single person to only a few items, can be fed back into the computer for a periodical update of the item parameter estimates.

An area of application where flexibility in item selection is most welcome is individualized instruction, for example, as implemented in Computer-Aided Instruction (CAI) systems. In such systems, the individual students' achievements are monitored by testing them regularly using short tests. The fact that the test

scores are used for instructional decision making entails the necessity of a high precision of measurement on a scale that is the same for all items.

From an optimization point of view, two different stages in item banking can be considered. The first is the calibration stage in which response data are collected with the primary intention to estimate the item parameters accurately (and to assess the fit of the items to the response model). Two obvious strategies in this stage are to collect the minimal number of item responses needed to guarantee a certain level of accuracy for the estimators, or to maximize the accuracy for a fixed number of responses. These two strategies can be implemented by manipulating the sampling design governing the collection of response data (see below). The second stage in item banking is the measurement stage. In this stage the items have already been calibrated and their parameters can, up to a tolerable inaccuracy, be considered as known. Now the optimization problem is to construct tests from the bank measuring the expected distribution of the persons' abilities with prescribed precision. Procedures for this optimization problem can be derived from linear programming (Boekkooi-Timminga, 1986; Theunissen, 1985; van der Linden & Boekkooi-Timminga, 1986).

The present paper deals with optimization of the item calibration stage. Typically, the design of a calibration sample has two characteristics: it is (1) structurally incomplete and (b) sequential by nature. The former property is a consequence of the fact that for item banks exceeding the size of a conventional test

it is physically impossible to administer all items to each subject in the sample. Hence, when designing the calibration samples, the item bank constructor has to decide for which item-person combinations responses will be collected and for which not. The latter indicates that item responses may be collected in more than one stage, notably when the first sample of subjects yield some parameter estimates for which the estimated standard error of estimation is considered too large and additional subjects have to be sampled.

It is the purpose of this paper to introduce a procedure that optimizes the calibration of an item bank fully sequentially and is so simple it can easily be implemented on a microcomputer. The procedure can even be applied at the level of pairs of item responses; then it automatically selects each next pair of items to be administered guaranteeing any predetermined level of accuracy for the item parameter estimates with a minimum number of responses. An obvious area for this level of applications is item banking in CAI systems where for individual students connected to the system the responses to previous items can be used to select the next items to be administered. However, application at the level of any other numbers of items and examinees in the steps of the procedure is equally well possible. Basically, the procedure consists of an empirical Bayes approach to a reformulation of the Rasch model as a model for paired comparisons with ties. Since this reformulation is a unique property of the Rasch model, the procedure cannot be used without modification for the other item

response models in use for item banking. Before this, first the Rasch model for paired comparisons with ties is presented. Then an empirical Bayes approach to estimating the item parameters in this model is discussed. Next, the sequential procedure for item bank calibration is introduced, first at the level of a single individual responding to a pair of items and subsequently for any number of individuals and items. Possible specifications for the first prior distribution in the approach are given in the subsequent section. In the Discussion the issue of sequential optimization of parameter estimation in other item response models is taken up again.

A Model for Paired Comparisons with Ties

In item response theory, the usual level of modeling is the response of a single subject to a single item. For instance, in the Rasch (1960) model, the probability of a subject p with ability α_p responding correctly to a dichotomous item i with difficulty δ_i is formulated as

$$(1) \quad P(X_{pi}=1|\alpha_p, \delta_i) = \alpha_p(\alpha_p + \delta_i)^{-1},$$

where $\alpha_p \in (0, +\infty)$, $\delta_i \in (0, +\infty)$ and X_{pi} is the random variable indicating whether the response is correct ($X_{pi} = 1$) or incorrect ($X_{pi} = 0$). A different perspective, however, is to consider a subject responding to pair of items i and j . If the values of the

item parameters are unknown, the outcome of this experiment can be conceived of as the outcome of a paired comparisons experiment in which the subject "judges" the relative difficulties of the test items. Three different outcomes can be distinguished:

$\{X_{pi} > X_{pj}\}$	item i is correct and j not;
$\{X_{pi} < X_{pj}\}$	item j is correct and i not;
$\{X_{pi} = X_{pj}\}$	both items i and j are correct or incorrect.

The first outcome can be interpreted as a comparison showing item j likely to be more difficult than i, whereas the second outcome indicates the reverse. The last outcome represents a compound event in which the two elementary events $\{X_{pi} = 0, X_{pj} = 0\}$ and $\{X_{pi} = 1, X_{pj} = 1\}$ have been taken together because, from the point of view of a comparison between the difficulties of the items, either indicates that i and j are likely to be equally difficult. In the literature on paired comparisons such outcomes are known as ties (Bradley, 1976; Davidson, 1970; Glenn & David, 1960; Kousgaard, 1976; Rao & Kupper, 1967).

From the model in Equation 1, completed with the property of local independence, the probabilities of the three possible outcomes follow straightforwardly:

$$\begin{aligned}
 P(X_{pi} > X_{pj}) &= \alpha_p \delta_j [(\alpha_p + \delta_i)(\alpha_p + \delta_j)]^{-1} \\
 (2) \quad P(X_{pi} < X_{pj}) &= \alpha_p \delta_i [(\alpha_p + \delta_i)(\alpha_p + \delta_j)]^{-1} \\
 P(X_{pi} = X_{pj}) &= (\alpha_p^2 + \delta_i \delta_j) [(\alpha_p + \delta_i)(\alpha_p + \delta_j)]^{-1}.
 \end{aligned}$$

The model in Equation 2 has been studied thoroughly in van der Linden and Eggen (1986). An important feature of the model not noted in the literature hitherto is that if all possible comparisons between a vector of n responses are considered, some of them are dependent. For example, if a person p has responded to three items i , j and k and the outcome $\{x_{pi} > x_{pj}\}$ has occurred, then the outcome $\{x_{pi} < x_{pk}\}$ is impossible. In van der Linden and Eggen (1986) it is shown that, for a person with responses to n items exactly $\min\{t, n-t\}$ of the $n(n-1)/2$ possible comparisons are independent, $t = \sum_{i=1}^n x_{pi}$ being the number-correct score of the person. Algorithms to remove dependent outcomes from the data are given in the same paper. The relationship of Equation 2 to other models for paired comparisons with ties is examined in Eggen and van der Linden (1987). These authors indicate that one of the well-known models for ties in the literature, viz. the one proposed by Davidson (1970), can be considered a simple reparameterization of the Rasch model for paired comparisons in Equation 2.

The outcome $\{x_{pi} = x_{pj}\}$ does not contain any information on the relative difficulties of the items. Disregarding this outcome, which is the same as conditioning the model in Equation 2 on the event of a non-tie $\{x_{pi} \neq x_{pj}\}$, yields the Bradley-Terry model from the paired comparisons literature:

$$(3) \quad P(X_{pi} > X_{pj} | X_{pi} \neq X_{pj}) = \theta_{ij} = \delta_j(\delta_i + \delta_j)^{-1}$$

(Bradley & Terry, 1952). Thus, for persons responding to a series of test items, if the dependent outcomes and the ties are removed from the data, a model extensively studied in the literature is available to analyze the remaining comparisons and estimate the item parameters. It should be noted that the parameter θ_{ij} in this model is a monotonic transformation of $\delta_i \delta_j^{-1}$ and therefore can be considered as a measure of the difficulty of item i relative to j . An important property of Equation 3 is that, unlike the model in Equation 2, it does not contain the person parameter α_p any longer. Hence, in order to estimate the parameters θ_{ij} , δ_i or δ_j the responses from any sample of subjects may be pooled. The fact that Equation 3 can be derived from the Rasch model was already noted in Rasch (1960). Fischer (1974, 1981) has used this derivation to establish several properties of the Rasch model. Choppin (1968), in an early but not widely noted paper, pointed at the advantage of using Equation 3 instead of 1 for item banking purposes (but ignored the aforementioned problem of dependency among the comparisons); the interest in the paired-comparisons design in the present paper was mainly motivated by this publication.

Parameter Estimation

From Equation 3 it follows that for a pair of items the number of outcomes $\{X_{pi} > X_{pj}\}$ for a series of independent trials can be described as a Bernoulli experiment with success parameter θ_{ij} . So, if n_{ij} , $i, j = 1, \dots, n$, is the number of non-ties and A_{ij} the number of times $\{X_{pi} < X_{pj}\}$ is observed, then

$$(4) \quad A_{ij} \sim \text{Bin}(n_{ij}, \theta_{ij}).$$

Hence, for a set I of independent comparisons between the n items, a product-binomial model with the following likelihood function holds

$$(5) \quad L(\delta_1, \dots, \delta_n; a_{ij}, n_{ij}, i, j) \\ = \prod_I [\delta_j(\delta_i + \delta_j)^{-1}]^{a_{ij}} [\delta_i(\delta_i + \delta_j)^{-1}]^{a_{ji}}.$$

The following algorithm for solving the likelihood equations from Equation 5 for $(\delta_1, \dots, \delta_n)$ has been introduced independently by Zermelo (1929) and Ford (1957) in the paired comparisons literature:

$$\begin{aligned}
 (6) \quad \delta_i^{(k+1)} &= \sum_{j=1}^n a_{ij} \left[\sum_{j=1}^n n_{ij} (\delta_i^{(k)} + \delta_j^{(k)})^{-1} \right]^{-1} \\
 &\quad i = 1, \dots, n \text{ cyclically.} \\
 \delta_j^{(k+1)} &= \delta_j^{(k)}, \quad j \neq i,
 \end{aligned}$$

In Equation 6 the superscript k indicates the iteration step and, as before, a_{ij} is a realization of A_{ij} and $n_{ij} = a_{ij} + a_{ji}$. The data needed to solve the likelihood equations are given schematically in Figure 1.

Insert Figure 1 about here

Existence and Uniqueness of Estimates

A theorem by Ford (1957) and Zermelo (1929) shows a necessary and sufficient condition for the existence and uniqueness of a solution to the likelihood equations defined by Equation 5. The condition reads: In every possible partition of the set of items into two non-empty subsets, for some item i in the first set and some item j in the second one, the outcome $\{X_{pi} > X_{pj}\}$ has occurred for at least one value of p .

The theorem has an instructive interpretation in graph theory: If the items are mapped one-to-one on the n points of a directed graph with an arrow from i to j if and only if $\{X_{pi} < X_{pj}\}$ has occurred at least once, the condition amounts to the requirement

that the resulting graph be strongly connected. The same condition has been established by Fischer (1981) for conditional maximum likelihood estimation of the item parameters directly from Equation 1. The condition is very mild and commonly met by data sets occurring in practice.

A practical consequence of the Zermelo-Ford condition is that it is not necessary for the design matrix in Figure 1 to be balanced, i.e. $n_{ij} = n$ for every (i,j) . Therefore, the reformulation of the Rasch model given above deals with incomplete data in the original person-by-item matrix in an elegant fashion. When analyzing the data at the level of pairs of items, incomplete data simply return as unequal numbers of comparisons per pair and maximum-likelihood estimation is still possible as long as the Zermelo-Ford condition is met. Experiments with an unequal number of repetitions have a tradition in the paired comparisons literature dating back to Dijkstra (1960).

An Empirical Bayes Approach

As noted earlier, calibration designs have structural incompleteness and the possibility of sequential optimization as typical features. The analysis above shows that the paired-comparisons design in Figure 1 has the simple provision of unequal numbers of repetitions n_{ij} to deal with structurally incomplete data. Since the model in Equation 3 does not contain the ability parameter α_p any longer, data from different subjects may be pooled

together. As a consequence, data from any subject, even when responding to only two items from the bank, can be stored in the format of Figure 1 and used in the estimation of the item parameters. The fact that this permits an efficient use of response data in item banking was already noted in Choppin (1968). From the point of view of data storage, Figure 1 also shows an efficient format. For n items, irrespective of the number of subjects, only an $(n \times n)$ -matrix has to be stored containing the counts of the numbers of times items i and j were compared and i turned out to be easier. Item calibration data are often collected in a sequential fashion. A paired-comparisons design would be fully appropriate for item banking purposes if sequential optimization of the calibration procedure were possible. The following presents such a procedure based on an empirical Bayes approach to the paired-comparisons model in Equation 3.

As A_{ij} is binomially distributed with $\theta_{ij} = \delta_j(\delta_i + \delta_j)^{-1}$ as success parameter, the natural choice of a prior distribution for this parameter in the Bayesian framework is the beta distribution. A variable θ has a beta distribution if its density is given by

$$(7) \quad f(\theta) = C(a,b) \theta^{a-1} (1-\theta)^{b-1}, \quad 0 \leq \theta \leq 1; a, b > 0,$$

where $C(a,b) = \Gamma(a+b)(\Gamma(A)\Gamma(B))^{-1}$ is a constant depending on the well-known gamma functions. Some properties of a beta-distributed variable useful in this context are

$$(8) \quad E(\theta) = a(a+b)^{-1}$$

$$(9) \quad \text{Mode}(\theta) = (a-1)(a+b-2)^{-1}$$

$$(10) \quad \text{Var}(\theta) = ab(a+b+1)^{-1}(a+b)^{-2}.$$

(Novick & Jackson, 1974, chap. 5). In general, the choice of the beta distribution as a prior for the binomial model is motivated by the following two properties:

1. The beta distribution is very flexible. Dependent on the values for the parameters a and b , it can take many shapes on the interval $[0,1]$. Nevertheless, its density function contains only two parameters.
2. The beta distribution is the natural conjugate of the binomial model. This means that the family of beta distributions is closed under the application of Bayes' theorem and that the posterior distribution will always be beta again. In other words, the transition from a prior to a posterior distribution takes place only at the parametric level; if θ has $B(a,b)$ as a prior and the data consists of n observations containing k successes, then θ has $B(a+k, b+n-k)$ as a posterior. The fact that the parameters a and b in the prior have the same effect on the posterior as the statistics k and n , respectively, gives the information in the prior an instructive interpretation: It is equal to the information about θ in a hypothetical series of n Bernoulli experiments with k successes.

In an empirical Bayes approach (Robbins, 1956, 1964) the Bayesian model is applied sequentially and the prior distribution for the model parameter(s) at step $k + 1$ is based on the observations at steps $1, \dots, k$. The process is repeated until the posterior distribution meets some stopping rule, e.g., allows making a decision with a certain expected risk. In the following section an empirical Bayes procedure optimizing the calibration of an item bank is given. For a previously determined level of uncertainty about the item parameters the procedure automatically minimizes the number of item administrations. The procedure operates at the level of pairs of items. Roughly, it starts with a first prior for all success parameters θ_{ij} and computes the posterior each time a paired comparison is made. The posterior is used as the prior for the next observation. At each step the pair of items with the largest uncertainty about θ_{ij} is selected for administration. The process is repeated until for all pairs the posterior uncertainty is below the maximum specified in a stopping rule.

Optimal Calibration Procedure

As in the preceding sections, it is assumed that the Rasch model in Equation 1 holds for $i = 1, \dots, n$, so that for every pair of items $i, j = 1, \dots, n$, the comparisons not resulting in a tie can be conceived of as the outcome of a Bernoulli experiment with probability of success $\theta_{ij} = \delta_j(\delta_i + \delta_j)^{-1}$. For every θ_{ij} ,

$i, j = 1, \dots, n$, as a prior the beta distribution in Equation 7 with parameters a_{ij} and b_{ij} is assumed. The variance $a_{ij}b_{ij}(a_{ij}+b_{ij}+1)^{-1}(a_{ij}+b_{ij})^{-2}$ in the distribution of θ_{ij} in Equation 10 is adopted as a measure for the uncertainty about the value of the parameter for the relative item difficulty θ_{ij} . Now, administering the pair (i, j) once at step k , the prior $B(a_{ij}, b_{ij})$ for $\theta_{ij,k}$ will be replaced by one of the following posteriors:

$$(11) \quad \begin{array}{ll} \{X_{pi} > X_{pj}\}: & \theta_{ij,k+1} \sim B(a_{ij}+1, b_{ij}) \\ \{X_{pi} < X_{pj}\}: & \theta_{ij,k+1} \sim B(a_{ij}, b_{ij}+1) \\ \{X_{pi} = X_{pj}\}: & \theta_{ij,k+1} \sim B(a_{ij}, b_{ij}) . \end{array}$$

Thus the outcome of a comparison between i and j is equivalent to the observation whether one of the parameters of the prior distribution must be updated by a value one. The posterior uncertainty about θ_{ij} is obtained by the same simple update of the parameters in Equation 11.

An optimal calibration procedure based on an empirical Bayes approach runs as follows:

1. Specify a first prior for the relative item difficulties θ_{ij} , $i < j = 1, \dots, n$, in the bank, i.e., select starting values for the parameters a_{ij} and b_{ij} .
2. Administer the pair of items with maximum variance (Equation 10):

$$(12) \quad \max_{(i,j)} a_{ij} b_{ij} (a_{ij} + b_{ij} + 1)^{-1} (a_{ij} + b_{ij})^{-2} .$$

If this pair is not unique, select one of the pairs satisfying

(12) at random.

3. Update the prior distribution of θ_{ij} according to Equation 11.
4. Repeat steps 2-3 until the maximum variance in Equation 12 is below a prespecified level.

It should be noted that this procedure can easily be implemented on a computer. Only a few lines of computer programming are needed and the data to be stored have already been displayed in the matrix in Figure 1. The parameters of the beta distribution are simply a_{ij} , the number of times i turns out to be easier than j , and $b_{ij} = n_{ij} - a_{ij}$, the remaining number of times i and j are compared, in this matrix. The choice of the first priors for the model amounts to an initialization of entries in the matrix, and the priors are simply updated by augmenting the entry a_{ij} by one if i was easier than j , b_{ij} by one if i was harder than j and neither if a tie occurred. The variance used in Equation 12 can be computed directly from these entries. The determination which pair of items has to be administered at each step can be simplified by storing the data in an array ordered by the size of the variances so that at each step the next pair to be administered is (a random selection of one of) the final element(s) in the array. The reason that the computations are not too intricate is due to the fact that no maximum-likelihood estimation of the individual item

parameters δ_i is needed in each step. The actual estimation of the item parameters is separated from the procedure optimizing its sampling design, the latter being based on the uncertainty about the relative item difficulties rather than their absolute values. However, at the same time the data needed for conditional maximum-likelihood estimation are kept in an appropriate format and each time actual estimates of the item parameters are needed the application of the Zermelo-Ford algorithm in Equation 6 to the data matrix in Figure 1 is straightforward.

As already mentioned, a natural environment to apply the above procedure at the level of pairs of items is item banking in computer-aided instruction. If an item bank in such systems has to be calibrated, the sampling of item responses can be adapted sequentially to the posterior uncertainties about the item parameters. If the calibration is finished, the direction of adaptation can be reversed to tailor the selection of items to the individual students' abilities.

Preposterior Analysis

A usual principle in sequential Bayesian decision making is preposterior analysis. For the present problem, preposterior analysis would mean that the selection of the pair of items at each step k is not based on the prior variances of θ_{ij} , $i < j = 1, \dots, n$, but on the expected reductions of the variances due to the outcome of the next comparison between the items. An obvious criterion

would be to choose the pair of items for which this reduction is maximal:

$$(13) \quad \max_{(i,j)} \phi(a_{ij}, b_{ij}) \equiv \max_{(i,j)} \{E[\text{Var}(\theta_{ij,k}) - \text{Var}(\theta_{ij,k+1})]\}$$

It is easily shown that this expected reduction is always positive, which is in agreement with our intuition that additional observations should reduce the uncertainty about parameter values. Although preposterior analysis is in the Bayesian tradition, it offers no advantage over Equation 12 in our case, the reason being that in the beta-binomial model equal prior variances imply equal expected reductions in variance.

Suppose the prior variances of $\theta_{i_1 j_1}$ and $\theta_{i_2 j_2}$ are equal:

$$(14) \quad \text{Var}(\theta_{i_1 j_1}) = \text{Var}(\theta_{i_2 j_2}).$$

Then,

$$\begin{aligned} & a_{i_1 j_1} b_{i_1 j_1} (a_{i_1 j_1} + b_{i_1 j_1} + 1)^{-1} (a_{i_1 j_1} + b_{i_1 j_1})^{-2} \\ &= a_{i_2 j_2} b_{i_2 j_2} (a_{i_2 j_2} + b_{i_2 j_2} + 1)^{-1} (a_{i_2 j_2} + b_{i_2 j_2})^{-2}. \end{aligned}$$

This holds only if

$$a_{i_1 j_1} b_{i_1 j_1} = a_{i_2 j_2} b_{i_2 j_2}$$

and

$$a_{i_1 j_1} + b_{i_1 j_1} = a_{i_2 j_2} + b_{i_2 j_2}.$$

Or, if

$$(15) \quad a_{i_1 j_1} = a_{i_2 j_2} \quad \text{and} \quad b_{i_1 j_1} = b_{i_2 j_2}$$

or

$$(16) \quad a_{i_1 j_1} = b_{i_2 j_2} \quad \text{and} \quad a_{i_2 j_2} = b_{i_1 j_1}.$$

Thus, Equation 14 implies Equation 15 or 16. Therefore, because

$$\phi(a,b) = \phi(b,a),$$

Equation 14 also implies equal reductions of the variances in Equation 13.

Since in the present case preposterior analysis does not introduce a more effective sampling of the item responses and Equation 13 is computationally more involved than Equation 12, the latter is the favorite criterion.

Generalization to Tests and Groups of Any Size

Thus far, at each step in the optimization procedure only responses

of a single subject to a pair of items have been considered.

Generalizations to tests and groups of subjects of any size are, however, straightforward.

The generalization for a single subject to a test of any length runs as follows: (1) Select a pair of items, say $i_0 j_0$, according to the above procedure, (2) suppress i_0 and j_0 in the array with prior variances and select the next pair, and (3) continue until enough items have been selected to fill the test.

A new aspect in this generalization is the independence of the outcomes of the comparisons necessary for using the likelihood equations from Equation 5. In principle, two strategies are available to satisfy the condition of independence. The first is to partition the test into pairs before the administration and to consider only the outcomes of the comparisons within these pairs. In this way independence is obtained by design. The other strategy is to use the fact that the number of independent non-ties in nonoverlapping pairs of items for a single subject and a test of n items is equal to $\min\{t, n-t\}$, and to select this number at random after the administration of the test under the restriction that for each subject an item figures only once in a pair. This strategy has been followed in van der Linden and Eggen (1986) and is recommended because it guarantees the maximum amount of data from the test.

The generalization to a test of any length for a group of subjects of any size poses no further problems. The same test is administered to all subjects. The fact that the data matrix is filled less efficiently than when the test for each next subject is

selected taking the responses of the previous subjects into account is the price one has to pay for leaving the domain of sequential optimization.

Specifying the First Priors

Four different ways of specifying the $\frac{1}{2} n(n-1)$ first priors in the procedure are possible. The first two are empirical methods; the other two subjective ones. Not all of these methods are feasible with larger item banks, though.

Empirical Information from Other Experiments

If the items were administered earlier in an experiment, the data from this experiment can be edited into the format of the matrix with paired comparisons in Figure 1; the entries of this matrix are the first values of the parameters of the priors in the procedure. In this approach again, in order to guarantee independent data either the pairs must be chosen before inspecting the data or $\min\{t, n-t\}$ comparisons must be sampled at random.

It should be noted that this method can be used even if not all of the items have been administered before. In that case, the priors for the other items have to be specified using one of the methods below. Again, this is an example of the efficient pooling of data from different sources possible because the model in Equation 3 is free from the person parameters α_p .

Empirical Information from the Experiment Itself

An obvious variant of the above method is to begin the calibration experiment without any prior and collect some data to fill the matrix in Figure 1 with starting values. For instance, before starting the empirical Bayes procedure all item pairs could be selected a few times in a random order to get empirical values for the parameters a_{ij} , $i < j = 1, \dots, n$. This approach might have appeal to those not so much in favor of using subjective priors.

Noninformative Priors

The beta distribution with parameter values $a = b = 1$ is the uniform distribution on the interval $[0, 1]$. The attractive thing about the choice of this distribution as the first prior for all item pairs is the fact that it does not favor any of the items; each item gets one "success" and one "failure" against every other item. Moreover, since the number of hypothetical comparisons ($n=2$) is small compared to the amount of data usually collected in an item calibration study, the effect of the subjectivity of this choice on the eventual estimates can be neglected. A final advantage of this method is that right from the beginning the data matrix meets the Zermelo-Ford condition for the existence and uniqueness of maximum-likelihood estimates; an adjacency matrix with positive (integer) values for its elements is strongly connected (Harary, Norman & Cartwright, 1965, theorem 5.14). As the

property of connectedness is maintained when data are added to the matrix, unique maximum-likelihood estimates always exist with uniform priors.

Subjective Priors

The final possibility is to have human judges specify subjective priors. An obvious procedure is to use a thought experiment based on the interpretation of the parameters a_{ij} in the beta distribution as the number of successes of item i against j in a hypothetical series of n_{ij} Bernoulli trials. In the two steps of this experiment the judge first has to estimate the percentage of times i will be easier than j , $a_{ij}n_{ij}^{-1} \times 100\%$, and then to indicate the number of empirical trials, n_{ij} , to which his/her certainty is equivalent. The subjective estimates could be smoothed by calculating the maximum-likelihood estimates of δ_i , $i = 1, \dots, n$, using the Zermelo-Ford algorithm and replacing the original subjective values for a_{ij} , $i < j = 1, \dots, n$, by their estimated expected values under the model. If desired, the residuals could, as is usual in paired-comparison experiments, be used to check the judges for inconsistencies in their judgments.

A problem with this method of subjective priors is that it cannot be used for larger numbers of items because then the number of comparisons soon gets forbiddingly large. However, the method can be used in combination with one of the above methods, for example, to fill the holes in the data matrix left after data from

previous experiments were used to get empirical values for the other elements.

Discussion

In the procedure of optimal item bank calibration presented in this paper the calculations necessary to decide which items to administer next were separated from the actual maximum likelihood estimation of the item parameters. Hence, a procedure demanding only a few simple calculations was possible. Nevertheless, the data are stored in a format for which an efficient algorithm for conditional maximum-likelihood estimation is available. An advantage of this procedure for conditional maximum-likelihood estimation from paired comparisons data is that, as opposed to regular conditional maximum-likelihood estimation, no numerical problems due to the presence of elementary symmetric function of an order equal to the test length occur (Gustafsson, 1980). Therefore, the procedure does not have the usual limitation to 50 or 60 items of the standard computer programs for the Rasch model and outperforms these considerably in required computing time (Eggen & van der Linden, 1987; van der Linden & Eggen, 1986).

A model for paired comparisons between test items without any person parameter, like the one in Equation 3, is only possible for item response models with a sufficient statistic for the person parameter. Among the logistic models with unknown item parameters this is a unique property of the Rasch model. Nevertheless, one may

wonder whether the same sequential optimization procedure is possible in attendance of, e.g., a discriminating power parameter in the model. For instance, a possible way out could seem to use double data storage, in the paired comparisons format to decide which items to administer next and in the usual (incomplete) persons $\times$ items format for item parameter estimation. However, even if maximization of the information about the difficulty parameter were possible along these lines, it still is the question whether this would imply simultaneous maximization of the information about the discrimination parameter. The point is that with more than one item parameter in the model the optimization problem becomes an example of decision making with multiple objectives. A possible solution to the problem of having more than one item parameter each imposing a different requirement on the sampling design adopted in van der Linden (1986) is to introduce a maximin criterion in the optimization procedure. Then it is possible to calculate an optimal sampling design for simultaneous administration of the test items to the subjects. The possibility of sequential procedures along these lines have to be investigated yet.

References

- Boekkooi-Timminga, E. (1986). Simultaneous test construction by zero-one programming. Methodika, 1, to appear.
- Bradley, R.A. (1976). Science, statistics and paired comparisons. Biometrics, 32, 213-232.
- Choppin, B.H.L. (198). An item bank using sample-free calibration. Nature, 219, 870-872.
- Davidson, R.R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. Journal of the American Statistical Association, 65, 317-328.
- Dijkstra, O. (1960). Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetition on pairs. Biometrics, 16, 176-188.
- Eggen, T.J.H.M., & van der Linden, W.J. (1987). The use of models for paired comparisons with ties. In W.E. Saris & I.N. Gallhofer (Eds.). Sociometric research: Volume 1. Data collection and scaling. London: Macmillan.
- Fischer, G.H. (1974). Einführung in die Theorie psychologischer Tests [Introduction to the theory of psychological tests]. Bern: Verlag Hans Huber.
- Fischer, G.H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. Psychometrika, 46, 59-77.

- Ford, L.R. Jr. (1957). Solution of a ranking problem for binary comparisons. American Mathematical Monthly, 64, 28-33.
- Glenn, W.A., & David, H.A. (1960). Ties in paired comparisons experiments using a modified Thurstone-Mosteller method. Biometrics, 16, 86-109.
- Gridgeman, N.T. (1959). Pair comparison, with and without ties. Biometrics, 15, 382-388.
- Gustafsson, J.E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. Educational and Psychological Measurement, 40, 377-385.
- Harary, F., Norman, R.Z., & Cartwright, D. (1965). Structural models: An introduction to the theory of directed graphs. New York: Wiley.
- Kousgaard, N. (1976). Models for paired comparisons with ties. Scandinavian Journal of Statistics, 3, 1-14.
- Novick, M.R., & Jackson, P.H. (1974). Statistical methods for educational and psychological research. New York: McGraw-Hill.
- Rao, P.V., & Kupper, L.L. (1967). Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. Journal of the American Statistical Association, 62, 194-204, corrigenda 63, 1550.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut.

- Robbins, H. (1956). An empirical Bayes approach to statistics. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1). Berkeley, California: University of California Press.
- Robbins, H. (1964). The empirical Bayes approach to statistical decisions problems. Annals of Mathematical Statistics, 35, 1-20.
- Theunissen, T.J.J.M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.
- van der Linden, W.J. (1986). Optimal sampling design for parameter estimation in item response models. Paper presented at the Annual Meeting of the Psychometric Society, Toronto, June 21-24.
- van der Linden, W.J., & Boekkooi-Timminga, E. (1986). Algorithmic test design with practical constraints. Enschede, The Netherlands: University of Twente, Department of Education.
- van der Linden, W.J., & Eggen, T.J.H.M. (1986). The Rasch model as a model for paired comparisons with an individual tie parameter. Enschede, The Netherlands: University of Twente, Department of Education.
- Zermelo, E. (1929). Die Berechnung der Turnierergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung [The computation of tournament results as a maximization problem in probability calculus]. Mathematische Zeitschrift, 29, 436-460.

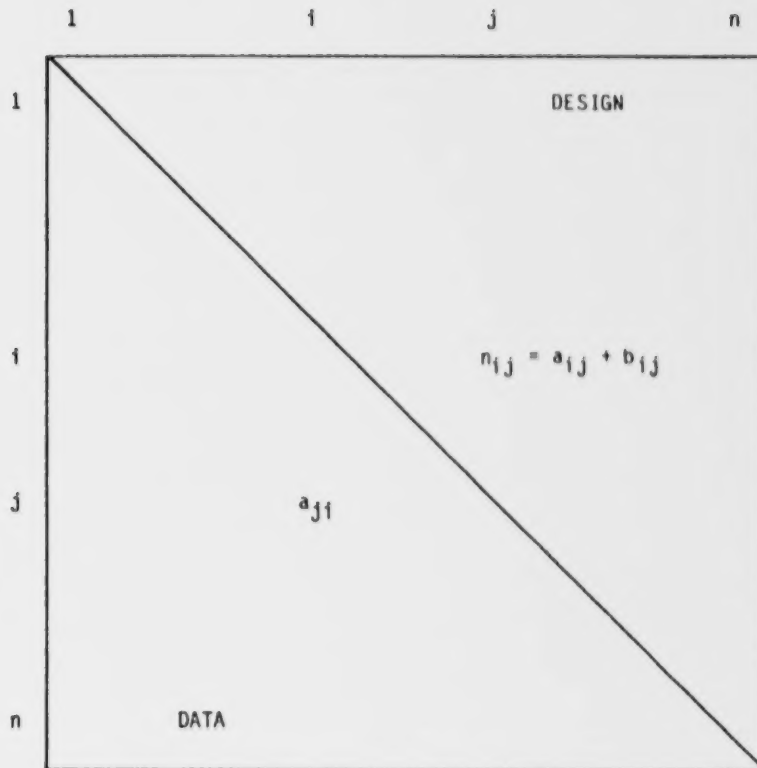


Figure 1. A paired-comparisons design for item banking

Notes

1. Now at the National Institute of Educational Measurement,
Arnhem, The Netherlands.

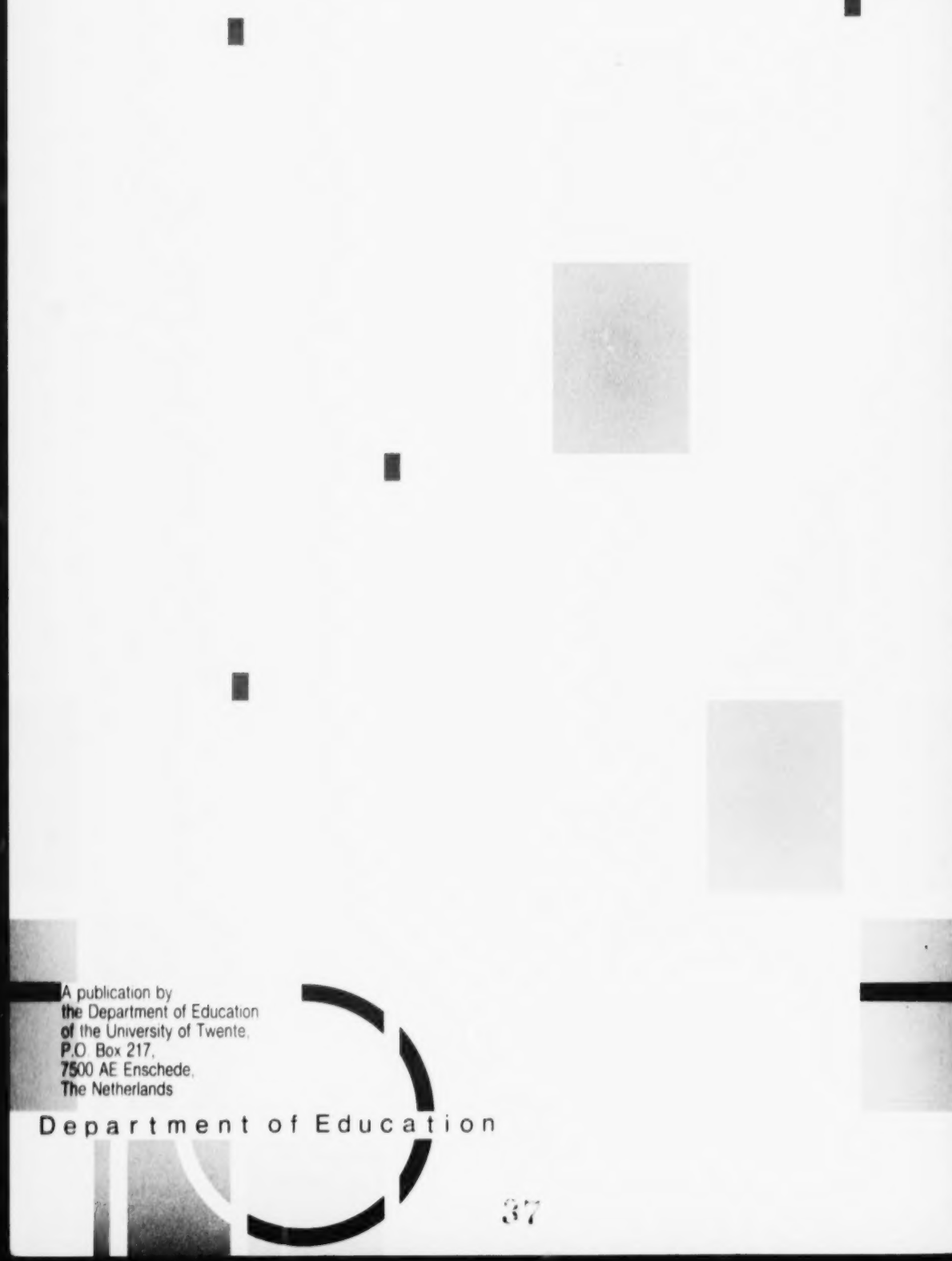
Authors' Notes

Request for reprints should be send to Wim J. van der Linden,
Universiteit Twente, Faculteit der Toegepaste Onderwijskunde,
Postbus 217, 7500 AE Enschede, The Netherlands or to Theo J.H.M.
Eggen, CITO, Postbus 1034, 6801 MG Arnhem, The Netherlands.

Titles of Recent Research Reports

- RR-86-1 W.J. van der Linden, The use of test scores for classification decisions with threshold utility
- RR-86-2 H. Kelderman, Item bias detection using the loglinear Rasch model: Observed and unobserved subgroups
- RR-86-3 E. Boekkooi-Timminga, Simultaneous test construction by zero-one programming
- RR-86-4 W.J. van der Linden, & E. Boekkooi-Timminga, A zero-one programming approach to Gulliksen's random matched subtests method
- RR-86-5 E. van der Burg, J. de Leeuw, & R. Verdegaal, Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features
- RR-86-6 W.J. van der Linden, & T.J.H.M. Eggen, An empirical Bayes approach to item banking
- RR-86-7 E. Boekkooi-Timminga, Algorithms for the construction of parallel tests by zero-one programming
- RR-86-8 T.J.H.M. Eggen, & W.J. van der Linden, The use of models for paired comparisons with ties

Research Reports can be obtained at costs from Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

The page features an abstract geometric design consisting of several squares and circles of varying sizes and shades of gray. Some shapes are solid, while others are outlines. These elements are scattered across the page, with a concentration of larger shapes near the bottom left and right edges. The overall aesthetic is minimalist and modern.

A publication by
the Department of Education
of the University of Twente,
P.O. Box 217,
7500 AE Enschede,
The Netherlands

Department of Education